

Reliability Estimation for a Software System with Sequential Independent Reviews

Nancy E. Rallis and Zachary F. Lansdowne

Abstract—Suppose that several sequential test and correction cycles have been completed for the purpose of improving the reliability of a given software system. One way to quantify the success of these efforts is to estimate the probability that all faults are found by the end of the last cycle. We describe how to evaluate this probability both prior to and after observing the numbers of faults detected in each cycle and we show when these two evaluations would be the same.

Index Terms—Bayes, confidence, estimation, software reliability.

1 INTRODUCTION

WE use the term “review” to refer to a software test and correction cycle. Suppose that a given software system is thought to contain faults and that a number of sequential independent reviews are conducted for the purpose of detecting those faults. These reviews are sequential in the sense that a fault detected during one review is corrected before the next review begins, so that the same fault cannot be detected during more than one review. One way to quantify the success of these efforts is to evaluate the probability that all faults are detected by the end of the last review and it is our measure of software reliability. We describe a Bayesian method for evaluating this probability based on the number of faults detected in each completed review cycle.

The sequential review model is related to several earlier treatments of software reliability. The time horizon for a testing process is sometimes divided into nonoverlapping intervals; in which case, those intervals could be regarded as a series of sequential reviews [6], [10], [11]. Our basic method of analysis, Bayesian inference, has been applied in many earlier studies on software reliability, e.g., [1], [2], [3], [7], [8], [9]. Those studies typically assume that observations of failure times are available. Our approach, however, does not use that kind of data but, instead, is based on the numbers of faults detected in a series of completed test and correction cycles.

The sequential model should be contrasted with the parallel review model, in which faults detected during one review could also be detected during any other review. In the “two-debugger estimation procedure,” for example, two debuggers work independently on the same program and could detect many of the same faults [13]. When there are parallel independent reviews, the probability that all faults

have been detected can be evaluated with capture-recapture sampling [12], [14].

For an example with sequential reviews, consider the year 2000 date problem. Many computer systems would have been adversely affected by the onset of the year 2000, unless actions were taken to review, test, and correct those systems [4]. The most common problem is the widespread use of two digits to represent a year, resulting in an inability to distinguish the correct century. To resolve this problem, the Department of Defense often required a software system to undergo five sequential reviews:

1. Identify faults based on notices from vendors or the Internet.
2. Identify faults based on reviewing available manuals.
3. Identify faults based on examining the source code or applying special tools.
4. Perform an “intrasystem” test by executing the isolated system in a post year 2000 environment.
5. Perform an “intersystem” test by executing the system as part of a network of systems in a post year 2000 environment.

Section 2 describes how to evaluate the probability that no undetected faults remain after a series of sequential independent reviews have been completed. Section 3 considers the validity of the underlying assumptions.

2 SEQUENTIAL INDEPENDENT REVIEWS

We wish to estimate the confidence that all faults have been detected in a system after several independent software reviews are completed. Let the index i refer to the i th review. We assume that these reviews are sequential and that they are ordered in such a way that review i is completed before review $i + 1$ begins. We also assume that all faults detected during one review are corrected before the next review begins, but without injecting any new faults during the correction process. The following notation is used:

Q_i = probability of detecting a given fault during the i th review.

• N.E. Rallis is with the Mathematics Department, Boston College, Chestnut Hill, MA 02167. E-mail: rallis@bc.edu.

• Z.F. Lansdowne is with the MITRE Corporation, 202 Burlington Rd., Bedford, MA 01730-1420. E-mail: zack@mitre.org.

Manuscript received 22 Jan. 1999; revised 25 Oct. 1999; accepted 16 Oct. 2000.

Recommended for acceptance by D. Hamlet.

For information on obtaining reprints of this article, please send e-mail to: tse@computer.org, and reference IEEECS Log Number 109022.

N_i = number of faults detected in the system during the i th review.

$E_i = \{N_1, \dots, N_i\}$

$p(K)$ = prior probability that the number of faults in the system is K before observing any data for the system.

$P_i(K|E_i)$ = conditional probability that K faults remain undetected in the system after i reviews are completed, given that the data E_i has been observed.

$P_i(K)$ = unconditional probability that K faults remain undetected in the system after i reviews are completed..

The detection probability Q_i is assumed to be the same for all faults. N_i is the number of faults detected in the i th review, and E_i is the set of the numbers of faults detected in the first i reviews; for $i = 0$, $N_0 = 0 = E_0$, and for $i = 1$, $E_1 = N_1$. The conditional probability $P_i(0|E_i)$ is the probability that no undetected fault remains, given that i reviews have been conducted and the data $E_i = (N_1, N_2, N_3, \dots, N_i)$ have been observed. The unconditional probability $P_i(0)$ can be interpreted as the probability that no undetected fault would remain if i reviews are conducted in the future. We are concerned with both the conditional and unconditional probabilities. In general, $P_i(0|E_i)$ is different from $P_i(0)$, but we will show when they are the same.

Define $B(m; p, f)$ to be the probability that m is the number of successes in f independent trials in which the probability of success at each trial is p . This probability is the well-known binomial distribution, and it can be written as

$$B(m; f, p) = \frac{f!}{(f-m)!m!} p^m (1-p)^{f-m}.$$

During any review, we assume that the process of detecting faults can be represented as a series of independent trials, one for each fault, and that each fault has the same probability of being detected. If K faults are present at the beginning of a review, and if Q_i is the probability of detecting each fault, then $B(N_i; K, Q_i)$ is the probability that exactly N_i faults are detected during that review.

With respect to the i th review, we may regard $P_{i-1}(K|E_{i-1})$ as being the prior distribution for K and $B(N_i; K, Q_i)$ as being the likelihood that N_i faults are detected. Thus, Bayes' Theorem enables us to evaluate the posterior probability with the following recursive formula [15]:

$$P_i(K|E_i) = \frac{P_{i-1}(K + N_i|E_{i-1})B(N_i; K + N_i, Q_i)}{\sum_{j=0}^{\infty} P_{i-1}(j + N_i|E_{i-1})B(N_i; j + N_i, Q_i)}, \quad (1)$$

where

$$P_0(K|E_0) = p(K). \quad (2)$$

This formula is appropriate because the posterior distribution from any given inspection becomes the prior distribution with respect to the next inspection, and Bayes' Theorem is applied to the next inspection's result.

The posterior probability can be evaluated numerically with (1) and (2) for any arbitrary distribution for the prior probability. In the special case in which the prior probability has the Poisson distribution, the following simplified result is obtained.

Theorem 1. *If the prior probability $p(K)$ has the Poisson distribution with mean λ_0 , then the posterior probability $P_i(K|E_i)$ has the Poisson distribution*

$$P_i(K|E_i) = \frac{(\lambda_i)^K}{K!} e^{-\lambda_i} \quad (3)$$

with mean

$$\lambda_i = \lambda_0 \prod_{j=1}^i (1 - Q_j). \quad (4)$$

Proof. The proof is by induction. By hypothesis, the prior probability is given by

$$p(K) = \frac{(\lambda_0)^K}{K!} e^{-\lambda_0}.$$

According to (1) and (2),

$$P_1(K|E_1) = \frac{p(K + N_1)B(N_1; K + N_1, Q_1)}{\sum_{j=0}^{\infty} p(j + N_1)B(N_1; j + N_1, Q_1)}.$$

After the expressions for $B(N_1; K + N_1, Q_1)$ and $p(K + N_1)$ are substituted into this formula, the terms $e^{-\lambda_0}$, $(K + N_1)!$, $(j + N_1)!$, $N_1!$, $Q_1^{N_1}$ and $\lambda_0^{N_1}$ can all be canceled, leaving

$$P_1(K|E_1) = \frac{\lambda_0^K (1 - Q_1)^K / K!}{\sum_{j=0}^{\infty} \lambda_0^j (1 - Q_1)^j / j!}.$$

Because the denominator is equal to $e^{\lambda_0(1-Q_1)}$, it follows that

$$P_1(K|E_1) = \frac{(\lambda_0(1 - Q_1))^K}{K!} e^{-\lambda_0(1-Q_1)},$$

which is the Poisson distribution with mean $\lambda_0(1 - Q_1)$.

Next, suppose that $P_{i-1}(K|E_{i-1})$ has been shown to be Poisson with mean $\lambda_0 \prod_{j=1}^{i-1} (1 - Q_j)$. Then, the posterior probability $P_i(K|E_i)$ is obtained by once again applying (1), where the Poisson distribution $P_{i-1}(K|E_{i-1})$ is the prior distribution. After substituting the formula for $B(N_i; K + N_i, Q_i)$ into (1) and simplifying, we obtain (3), where λ_i satisfies (4). Thus, by induction, formulas (3) and (4) hold for any number of reviews i . \square

Corollary 1. *If the prior probability $p(K)$ has the Poisson distribution with mean λ_0 , then the confidence that no faults remain after i independent reviews with observations E_i is*

$$P_i(0|E_i) = e^{-\lambda_i}, \quad (5)$$

where λ_i is given by (4).

Proof. Follows immediately from Theorem 1. \square

Theorem 2. For any prior distribution $p(K)$, the unconditional probability that no fault will remain after conducting i independent reviews can be evaluated as

$$P_i(0) = \sum_{K=0}^{\infty} p(K) \left(\sum_{j=1}^i q_j \right)^K, \quad (6)$$

where

$$q_j = Q_j \prod_{m=1}^{j-1} (1 - Q_m). \quad (7)$$

Proof. Note that q_j is the probability that a fault is detected for the first time during the j th review because this event can occur only if it was detected during the j th review and not detected during the preceding $j - 1$ reviews. Because the q_j refers to mutually exclusive events, namely, that a fault is detected during a particular review, $\sum_{j=1}^i q_j$ is the probability that a fault is detected sometime during i reviews. If the detection of each fault is independent of the detection of any other fault,

$$P_i(0|K) = \left(\sum_{j=1}^i q_j \right)^K$$

is the conditional probability that no fault remains after i reviews, given that the total number of faults is K . By the law of total probability

$$P_i(0) = \sum_K P_i(0|K)p(K)$$

which implies (6). \square

Theorem 3. If the prior probability $p(K)$ has the Poisson distribution with mean λ_0 , the unconditional probability that no fault will remain after conducting i independent reviews can be evaluated as

$$P_i(0) = e^{-\lambda_i}, \quad (8)$$

where λ_i is given by (4).

Proof. By Theorem 2,

$$\begin{aligned} P_i(0) &= \sum_{K=0}^{\infty} \frac{\lambda_0^K e^{-\lambda_0}}{K!} \left(\sum_{j=1}^i q_j \right)^K \\ &= e^{-\lambda_0} \sum_{K=0}^{\infty} \frac{1}{K!} \left(\lambda_0 \sum_{j=1}^i q_j \right)^K. \end{aligned}$$

The righthand side of the above formula is $e^{-\lambda_0}$ times the power series expansion of the exponential function with mean $\lambda_0 \left(\sum_{j=1}^i q_j \right)$ and, so,

$$P_i(0) = e^{-\lambda_0 \left(\sum_{j=1}^i q_j \right)}.$$

We can write

$$1 - \sum_{j=1}^i q_j = \prod_{j=1}^i (1 - Q_j)$$

because both sides represent the probability that a fault is not detected during the i reviews. Since λ_i is defined by (4), (8) holds. \square

If the prior probability has the Poisson distribution, then Corollary 1 and Theorem 3 show that the conditional and unconditional probabilities $P_i(0|E_i)$ and $P_i(0)$ are the same. In this case, the confidence that all faults have been found is a function of the number of reviews, the detection probabilities, and the mean of the prior distribution, but it is not a function of the numbers of faults actually observed during the successive reviews. This result seems remarkable because it gives a circumstance in which the statistical confidence from a Bayesian analysis is actually independent of all observed data.

When the observed data consists of times to failure, Landberg and Singpurwalla [8] derived results related to Theorem 1 and Corollary 1, and Musa et al. [10] derived results related to Theorem 3. Landberg and Singpurwalla showed that a Poisson prior distribution yields a Poisson posterior distribution that depends upon the observed failure times. In contrast, Theorem 1 and Corollary 1 show that the resulting Poisson posterior distribution is independent of observations when those observations are the numbers of faults detected in the review cycles. Musa et al. showed that a Poisson prior distribution combined with an exponential distribution for the time between failures yields a Poisson distribution for the number of failures remaining at any given time. In contrast, Theorem 3 shows that an exponential formula could be used to evaluate the probability that no fault remains when a Poisson prior distribution is combined with a binomial detection process in each review cycle.

The foregoing theorems enable the following question to be answered under varying circumstances: How many independent reviews are needed to achieve a desired confidence that all faults have been found in a given software system? If the prior probability can be represented with the Poisson distribution, Corollary 1 and Theorem 3 show that this question has the same answer whether or not any reviews have been conducted and any data have been collected. Suppose that the prior probability cannot be represented with the Poisson distribution. If it is possible to schedule an additional review after completing any number of reviews, then (1) and (2) can be used to determine whether such an additional review is needed to achieve the desired confidence. However, suppose that the budgeting and planning cycles are such that all reviews must be scheduled prior to conducting any of them. In the latter case, Theorem 2 can be used to determine the number of reviews that are needed.

These theorems do not require all software reviews to be identical because the detection probability Q_i can vary from review to review. Reviews with higher detection probabilities are more effective than those with lower detection probabilities.

Fig. 1 illustrates the application of Corollary 1 and gives the posterior probability of not having any fault under the following conditions: the prior distribution is Poisson with mean $\lambda_0 = 10$; the number of sequential reviews is either 2 (lowest curve), 3, 4, or 5 (highest curve); the detection

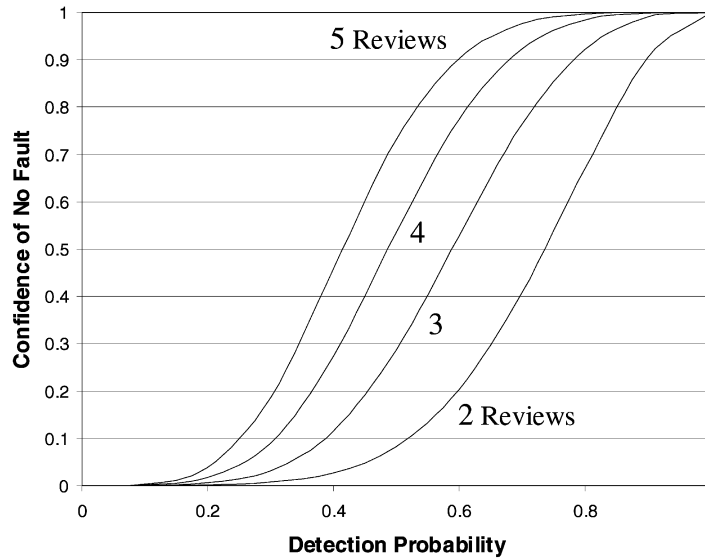


Fig. 1. Confidence of no remaining fault as a function of the detection probability and number of sequential reviews.

probability is the same for all reviews and varies from 0 to 1. For example, if the detection probability during each review is 0.6, this figure shows the confidence that no fault remains after i sequential reviews to be 0.20 for $i = 2$, 0.53 for $i = 3$, 0.77 for $i = 4$, and 0.90 for $i = 5$. On the other hand, suppose that we require the confidence of not having a fault to be at least 0.8. To achieve this confidence after i sequential reviews, the detection probability must be at least 0.85 when $i = 2$, 0.72 when $i = 3$, 0.62 when $i = 4$, and 0.53 when $i = 5$.

3 ASSUMPTIONS

The foregoing results are based upon a number of key assumptions. Let's consider the validity of those assumptions. The first key assumption is that each fault has the same probability of being detected during each review. This assumption is typically made for parallel review models that are analyzed by capture-recapture sampling [13], [14]. Also, time-to-failure models characteristically assume that the time to failure of a fault is identically distributed for all faults [10]. In reality, however, the various faults may vary in difficulty of finding them. For example, "easy" faults might be found with probability 0.9, while "difficult" faults might be found with probability 0.4. If there were multiple types of faults, it might be thought that a separate analysis could be performed for each type. Such an analysis, however, would require there to be a small, finite, known number of fault "types" and an ability to recognize in advance the type to which each fault belongs. These two conditions may be difficult to satisfy because it seems more likely that each fault has a unique probability of being found.

The second key assumption is that the detection can be treated as a series of independent trials with each fault corresponding to a trial. This assumption is also typically made for parallel review models [13], [14]. In Musa et al. [10], failures caused by faults are assumed to occur independently over time. The assumption of independent

trials does not seem to be appropriate for a test review that executes a software system with a realistic set of inputs for a specific duration of time; in this case, all faults might not be tested because the sequence of executed instructions might skip over portions of the code. To clarify this argument, consider two faults, A and B, that belong to the same segment of code and let $P(A)$ and $P(B)$ be the probabilities of finding A and B, respectively, on a randomly-chosen test run; then, $P(A|B) > P(A)$ because finding B implies a higher probability (than the marginal) of the executed test run being one that executes the specific segment that A and B reside on.

It might be thought that the assumption of independent trials would be appropriate for some types of reviews, such as code analysis. Here, we might think of a debugger who scans every line of source code and who recognizes, with a certain probability, that a given line with a fault has a fault. In reality, however, many faults may not be "in" any one line of code but are part of how the code is organized. Even if all faults were "in" individual lines of code, we could expect that a debugger's results would depend on learning typical or recurrent fault patterns within the code and, in such a case, possibly invalidate the assumption of independent trials.

The third key assumption is that all faults detected during a review are corrected prior to the next review and that new faults are not introduced during a correction process. This assumption can be relaxed and replaced by the more cumbersome formulation: Among the faults detected during a review, it is possible to identify any faults detected during earlier reviews but not corrected, and it is possible to identify any faults introduced during earlier correction processes. If we define N_i to be the number of faults detected in the i th review that were not previously detected and were not introduced by earlier correction processes, then the earlier theorems provide a way of estimating the number of original faults that remain undetected.

Theorems 1 and 3 and Corollary 1 assume that the prior probability for faults has the Poisson distribution. The software reliability literature (e.g., [8], [10]) sometimes uses this assumption as a springboard for either a classical or Bayesian analysis when estimating the number of remaining faults based on observed failure times. The Poisson distribution is unbounded, which means that it gives a positive probability for an arbitrarily large number of faults. When might this distribution be reasonable for the prior probability? The i th line of code could be thought of as having some prior probability p_i of containing a fault, where this probability may vary from line to line. In this case, the total number of faults S_n is the sum of n Bernoulli trials with variable probabilities, where n is the number of lines of code. If these trials are independent, if $\lambda_0 = p_1 + p_2 + \dots + p_n$ has a moderate value, and if n is large, then Feller [5, p. 264] showed that the distribution of S_n can be approximated by the Poisson distribution with mean λ_0 . In practice, we expect n to be several orders of magnitude larger than λ_0 . The only condition that may be doubtful is the independence assumption. Many faults might be due to misunderstandings of requirements or specifications. If a misunderstanding happens, which is a chance event, a large number of faults might be present in multiple lines of the code, thereby violating the independence assumption.

In conclusion, we have analyzed a sequential review case with what are perhaps the simplest and most tractable assumptions. Given those assumptions, we quantified the success of sequential testing and correction cycles using Bayesian methods. In particular, we evaluated the confidence that all faults in a system are detected by the end of a given cycle. We derived formulas for evaluating this confidence both prior to and after observing the number of faults detected in each cycle and showed under what condition these two measures are the same.

REFERENCES

- [1] M.-A. El-Aroui and J.-L. Soler, "A Bayes Nonparametric Framework for Software-Reliability Analysis," *IEEE Trans. Reliability*, vol. 45, no. 4, pp. 652-660, Dec. 1996.
- [2] G. Becker and L. Camarinopoulos, "A Bayesian Estimation Method for the Failure Rate of a Possibly Correct Program," *IEEE Trans. Software Eng.*, vol. 16, no. 11, pp. 1307-1310, Nov. 1990.
- [3] A. Csenki, "Bayes Predictive Analysis of a Fundamental Software Reliability Model," *IEEE Trans. Reliability*, vol. 39, no. 2, pp. 177-183, June 1990.
- [4] Department of Defense, *Year 2000 Management Plan*, Office of the Assistant Secretary of Defense (Command, Control, Communications, and Intelligence), Apr. 1997.
- [5] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. I, second ed. Wiley, 1957.
- [6] A.L. Goel, "Software Reliability Models: Assumptions, Limitations, and Applicability," *IEEE Trans. Software Eng.*, vol. 11, no. 12, pp. 1411-1423, Dec. 1985.
- [7] W.S. Jewell, "Bayesian Extensions to a Basic Model of Software Reliability," *IEEE Trans. Software Eng.*, vol. 11, no. 12, pp. 1464-1471, Dec. 1985.
- [8] N. Langberg and N. Singpurwalla, "A Unification of Some Software Reliability Models," *SIAM J. Scientific and Statistical Computing*, vol. 6, no. 3, pp. 781-790, July 1985.
- [9] T.A. Mazzuchi and R. Soyer, "A Bayes Empirical-Bayes Model for Software Reliability," *IEEE Trans. Reliability*, vol. 37, no. 2, pp. 248-254, June 1988.
- [10] J.D. Musa, A. Iannino, and K. Okumoto, *Software Reliability: Measurement, Prediction, and Application*. McGraw-Hill, 1990.
- [11] G.J. Schick and R.W. Wolverton, "An Analysis of Competing Software Reliability Models," *IEEE Trans. Software Eng.*, vol. 4, no. 2, pp. 104-119, Mar. 1978.
- [12] G.A.F. Seber, *The Estimation of Animal Abundance and Related Parameters*. Macmillan, 1982.
- [13] M.L. Shooman, *Software Engineering: Design, Reliability, and Management*. McGraw-Hill, 1983.
- [14] S.A. Vander Wiel and L.G. Votta, "Assessing Software Designs Using Capture-Recapture Methods," *IEEE Trans. Software Eng.*, vol. 19, no. 11, pp. 1045-1054, Nov. 1993.
- [15] R.L. Winkler and W.L. Hays, *Statistics: Probability, Inference, and Decision*, second ed. Holt, Rinehart, and Winston, 1975.



Nancy E. Rallis received the BA degree from Vassar College, and the MA and PhD degrees in mathematics from Indiana University. She is an associate professor of mathematics at Boston College, Chestnut Hill, Massachusetts.



Zachary Lansdowne received the BS and MS degrees in electrical engineering from the Massachusetts Institute of Technology, an MA degree in psychology from Antioch University in Los Angeles, and a PhD degree in operations research from Stanford University, Palo Alto, California. He is with the Economic and Decision Analysis Center at The MITRE Corporation, Bedford, Massachusetts. He held previous positions with the Rand Corporation in Santa Monica, California, and the Systems Optimization Laboratory at Stanford University.

► For more information on this or any computing topic, please visit our Digital Library at <http://computer.org/publications/dlib>.